ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

IJARCCE

nCORETech

LBS College of Engineering, Kasaragod

Vol. 5, Special Issue 1, Feburary 2016

Transliteration in Malayalam Using Deep Learning

Shilpa Krishnan[1], Usha K[2]

PG Student, Dept. of CSE, NSSCE, Palakkad, India [1]

Associate Professor, Dept. of CSE, NSSCE, Palakkad, India [2]

Abstract: Transliteration: The process of transcription from one language to another is an inevitable part of Machine Translation. It is concerned about Out-of-Vocabulary words or Named Entities and Proper Nouns which requires its phoneme or sound preservation across the language. It can be done in a supervised or unsupervised way. But there is no good method which does this in Malayalam language. Deep Learning is an emerging stream which can be used in the implementation of the same. This paper throws light into such a method which is used to implement Machine Transliteration in Malayalam using Deep Learning using unsupervised training method.

Keywords: Transliteration, Unsupervised Learning, Deep Learning, Deep Belief Nets (DBN), Restricted Boltzmann Machine (RBM).

I. INTRODUCTION

Natural Language Processing: a field to investigate how well a machine process its human master's language, is one of the most promising and highly research oriented field of Computer Science. In an attempt to realize this thought many Machine Translation Systems for different languages have been put forth. But once the focus is turned to regional languages like Hindi or Malayalam, though a lot of efforts are seen, a fully fledged system does not exist. So the field is really challenging and promising as there are large chances of vast improvement. Regional language MT systems still need many of its functionalities or modules to be modified or implemented using better methods.

Most of the MT systems in Indian language, like Malayalam are based on rule based methods which is apt for the behavior of such language. One of the most important but yet untouched field of MT system is transliteration. Transliteration is the translation of the sounds in one language to another language or it is the conversion of one language alphabet to another language alphabet by preserving the sound. Though a little work is seen in this field, all of them are done in a supervised way. The advent of deep learning, a part of neural networks has shone light to a new research direction in MT systems as well as in many fields of NLP giving better options and result. Using deep nets in deep learning technology actually opens to unsupervised learning methods which is a forward step to make the MT systems more brilliant.

The transliteration system which is a phonetic converter or a transcription system is based on deep net which involves an unsupervised method for training. Deep nets are based on the concepts of RBMs and neurons. Many layers are stacked one over the other to form multilayer RBMs with each layer having a number of neurons or nodes.

II. RELATED WORKS

Though many transliteration works across many different languages have been emerged, no much progression or work is seen in Malayalam language. The only paper so far traced in Malayalam language is "English to Malayalam Transliteration Using Sequence Labeling Approach" by Sumaja et. al.[4]. Here in this approach support vector method which is a supervised learning mode is used. SVMTool is used for transliteration purposes. The system has 3 phases which includes preprocessing, training phase and transliteration phase with main 2 steps of segmentation of the source string into transliteration units and mapping the source language units to target language unit.

Another paper which deals with transliteration is "Handling Unknown Words in Named Entity Recognition Using Transliteration" [5] by Deepti et. al. Here using an algorithm they extract the Named Entities (NER) and manually provide their transliteration units to different languages and save the words in the database with a tag attached to it. When new sentences are given they will check the units with NER tags with the already created database and see if corresponding transliterated pair is present in the database. If such pairs are present, then that corresponding unit is substituted else the word to be transliterated is added to the database with its corresponding transliteration units in the languages which are wanted. This is a very trivial approach which involves the maintaining of the database manually.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

IJARCCE

nCORETech

LBS College of Engineering, Kasaragod

Vol. 5, Special Issue 1, Feburary 2016

In "Hybrid Approach to English-Hindi Name Entity Transliteration" [2] the authors tried to identify 7 different phonemes for English words, which are group of consonants and vowels, by defining rules for capturing phonemes in the training phase. Then the NER tagged English words are extracted and their phonemes are found out by applying the phoneme extraction algorithm after which their corresponding Hindi phoneme is found out and saved to create a knowledge base along with the corresponding English phonemes. Here the ngram probability is used to calculate the probability of the English – Hindi phoneme knowledgebase here. In the testing phase when a new NER tagged word is encountered it is passed through the phonification module and its corresponding Hindi syllable with highest probability is taken and Hindi word is formed.

Another paper more related to the work going to be proposed is "A Deep Learning Approach to Machine Transliteration" [1] by Thomas et. al. which is a transliteration system between Arabic - English languages. It is based on Deep Belief Nets. Here 3 layers of RBMs are used each for source and the target language parts. A joint layer for the conversion of source to target language is kept as the top layer of the system.
A more close concept which is proposed in one of the Indian language is "Joint Layer-based Deep Learning Framework for Bilingual Machine Translation (For English and Tamil Languages)" [3] by Sanjanashree et. al. Here also the method opted is based on Deep Belief Nets. Here 2 layers of RBM are used each at source and target languages and a joint layer is implemented at the top of the system. Each layer has a large number of neurons. This is a much more related work to the system to be proposed both by the technology used and the similarity and characteristics of languages.

III.MOTIVATION

Malayalam is one of the classical Indian language. Many machine translations systems have been proposed in Malayalam language but no transliteration system other than the already mentioned SVM based system is proposed in Malayalam. So, as one of the important Indian languages it needs a good Transliteration system and that too an unsupervised one to club with the translation system.

IV.PROBLEM DEFINITION & PRELIMINARIES

Transliteration, the transcription of words such as proper noun from one language to another or more commonly from one alphabet to another is an important subtask of machine translation in order to obtain a high quality output. So the aim is to implement an unsupervised English to Malayalam transliterator based on Deep Belief Nets.
There are 3 types of transliteration: Grapheme based, Phonetic based, Correspondance based. In grapheme based transliteration each syllable is converted to its corresponding target language syllable whereas in phoneme based transliteration phonetic sound is maintained and the converted to target language. Correspondence based is the combination of grapheme based and phonetic based.
Deep Belief Networks are generative models based on RBMs, which have multiple layers in their architecture for extracting different features of data or objects presented to them, and follows an unsupervised learning method. They can easily handle unlabelled data. They can also extract the features required to model a classifier by themselves and work on it.
RBM (Restricted Boltzmann Machine), which is a descendent of the perceptron in artificial neural network family, is a bipartite undirected graphical models. Usually a basic RBM have 2 parts : a visible set of nodes and a hidden layer of nodes. There won't be any connection among the visible layer nodes and similar is the case with nodes in hidden layer. But undirected connection will be extended between the hidden layers nodes and visible layer nodes. There can be many hidden layer nodes at the top of visible layer nodes. The visible nodes take the input and usually are binary units. Only the visible unit values or input are known. The hidden unit or node learns the required relevant featur0es. An energy function is calculated for the joint assignment using the weights assigned as well as the values of visible and hidden nodes and is propagated to upper layer, if more than 2 layers are present for the RBMs, as the input for the next hidden layer.
The energy assignment for joint assignment can be given as follows:

$$E(v,h) = \sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij} v_i h_j + \sum_{i=1}^{n} b_i v_i + \sum_{j=1}^{m} c_j h_j$$

Where v_i and h_j represents the nodes at a visible layer and hidden layer respectively and w_{ij} represents the weight and b_i and c_j represents the bias vectors. The joint probability with partition function Z is given by the equation:

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

IJARCCE

nCORETech

LBS College of Engineering, Kasaragod

Vol. 5, Special Issue 1, Feburary 2016

$$P(v,h) = \frac{\exp^{-E(v,h)}}{Z}$$

The conditional probability of the visible unit 'v' given hidden unit 'h' and vice versa can be represented as:

$$P(v = 1|\, h) = \text{sigmoid}\,(b + \sum W * h)$$

$$P(h = 1|v) = \text{sigmoid}(c + \sum W * v)$$

A DBN is created by stacking the RBMs. The input is given as a binary matrix usually a sparse one and the first RBM is trained on it. After that a greedy layer wise training is undergone with the output of the first RBM layer as the input of the next layer until a satisfactory level of depth is reached. The layers learn data from the input raw data in an unsupervised mode and recreates or reconstructs the data effectively.

V. PROPOSED METHOD

The proposing method, which is an extension of [3], is based on DBN and is for English to Malayalam Transliteration. The system should have 2 parts: a source part which is trained on the source language and its characteristics, and a target part which is trained on the target language and its characteristics during the training stage. Each side will have two layers of RBM termed as source encoders and target encoders. Initially the source and target encoders are trained separately and independently. The encoders have a joint layer at the top as a third layer which does the actual mapping of source language features to target language features. So the joint layer acts as a bridge or connection between the encoders at the two sides. Each layer will have 'n' neurons and the same level at the source and target encoders will have same number of neurons. The number of neurons at each layer is directly proportional to the features learned or to the information extracted. So higher the number of neurons at each layer more will be the information extracted. But as the number of neurons increases the complexity of the system increases drastically and unaffordably, which may also affect the system speed.

The number of layers in the system is fixed as two, based on the similarity of the Malayalam language to the Tamil language, as both fall under the category of Dravidian languages. Usually two layers are chosen for the Deep Belief Nets as more number of layers will increase the complexity and two layers will be sufficiently good for the purpose. The number of neurons in each layer is decided based on experimentation with arbitrary number of neurons in each layer, followed by their training and testing the efficiency of output obtained after training. The number of neurons which give maximum efficient output is decided as the final systems configuration.

The working of the system can be traced as follows: First pre-processing of data like Romanization of Malayalam words should be done. Then the input in both source and target language is to be converted to binary representation which is fed as input to the source and target encoders separately. The output of first layer will serve as input to the second layer. Finally the outputs of both the encoders are fed to the joint layer for the mapping of two scripts. This is carried out for each named entity in the training phase. Each layer of DBN can be trained using Contrastive Divergence (CD) algorithm to optimize the weights and bias terms.

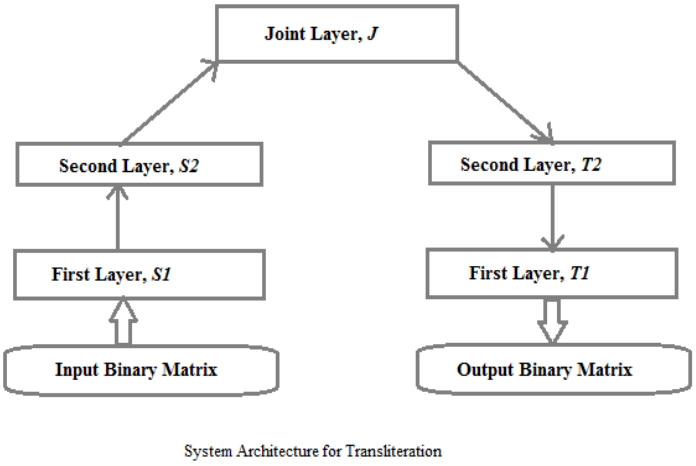

Fig. 1 System Architecture for Transliteration

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

IJARCCE

nCORETech

LBS College of Engineering, Kasaragod

Vol. 5, Special Issue 1, Feburary 2016

During the testing or transliterating phase, the input is given at the source visible layer. The 1st layer of the source input will calculate an output and it is taken as the input to the next layer. The output of the second layer is in turn given as the input to the joint layer. The joint layer will take this as input and gives an output to the second layer of the target encoder which in turn produces an output which is given as input for the first layer of target encoder and this down ward traversing takes place to the visible layer of target encoder which gives the output in target language. The system architecture is shown in figure 1.

VI. COMPARISON & EXPECTED RESULT

This proposed system is a theoretical one which is yet to be implemented. Through the studies conducted the expected result and its comparison with other systems is explained as below.

The system based on "Hybrid Approach for English-Hindi Transliteration System for Proper Nouns" follows a supervised learning mode and is based on statistical approach which gives an accuracy of 97% which is trained with 18000+ names and 90000+ ngrams.

The system based on "Hybrid Approach to English-Hindi Name Entity Transliteration" also follows an supervised learning method which is based on phoneme extraction and gives a result of 83% accuracy when trained with 42,371 name entities.

The English to Malayalam Transliteration Using Sequence Labeling Approach" follows supervised learning which uses the support vector machines and has 90% accuracy when trained with 20,000 names.

The proposed system follows an unsupervised mode whose underlying architecture is deep belief nets and is expected to give a result with 80% accuracy with around 3000-4000 names trained.

The study shows that with much less training data the system tends to show a good accuracy rate which is expected to increase as the number of training names increases.

VII. CONCLUSION & FUTURE WORKS

Deep Learning is a relatively new technology which can cause great leaps in the Natural Language Processing field. The main driving factor is the unsupervised way of feature learning of unlabelled data for transliteration. Here the system tries to reconstruct the data using learned features. It also has a better performance in case of accuracy and result.

The future works include the implementation of model and it can be extended to Malayalam to English transliteration as the RBM has bidirectional properties and undirected structure.

ACKNOWLEDGMENT

First and foremost I thank the almighty god for helping me to fulfill this work. I also extend my sincere thanks from the depth of my heart to my guide for enlightening me through her valuable guidance, help and suggestions. I would also like to extend my gratitude to all my friends, dears and nears for helping me to make this work a success.

REFERENCES

[1] Veerpal Kaur, Amandeep kaur Sarao, Jagtar Singh, "Hybrid Approach for English-Hindi Transliteration System for Proper Nouns", International Journal of Computer Science and Information Technologies, Vol. 5(5), pp 6361-6366,2014.
[2] Sruthi Mathur,Varun Prakash Saxena, "Hybrid Approach to English-Hindi Name Entity Transliteration."
[3] Sanjanashree P., Anand Kumar M,Joint Layer based Deep Learning Framework for Bilingual Machine Transliteration (For English And Tamil Languages), IEEE International Conference on Advances in Computing Communications & Informatics, 2014.
[4] Sumaja Sasidharan, Loganathan R, Soman K P, "English to Malayalam Transliteration Using Sequence Labeling Approach",International Journal of Trends in Engineering Vol. 1, No. 2, May 2009.
[5] Deepti Chopra, Sudha Morwal, Dr. G. N. Purohit "Handling Unknown Words in Named Entity Recognition Using Transliteration" International Journal on Natural Language Computing (IJNLC) Vol. 2, No.4, August 2013.
[6] Salakhutdinov, Hugo Larochelle "Efficient Learning Of Deep Boltmann Machine" 13th Internation Conference on Artificial Intelligence and Statistics (AISTATS) Vol. 9 JMLR 2010."